DEVELOPMENT OF CSLU LVCSR: THE 1997 DARPA HUB4 EVALUATION SYSTEM

Yonghong Yan   Xintian Wu   Johan Schalkwyk   Ron Cole

Center for Spoken Language Understanding
Oregon Graduate Institute of Science and Technology
P.O. Box 91000, Portland, OR 97291-1000
{yan, xintian, johans, cole}@cse.ogi.edu

ABSTRACT

This paper presents the CSLU Broadcast News transcription system used in the DARPA 1997 evaluation. The system was built using the softwares developed for the CSLU LVCSR project started in January 1997. This 25K-word vocabulary system used continuous HMMs for acoustic modeling and the standard backoff trigram as the language model. The search used a single pass decoder with MLLR based adaptation technique. Although on the standard DARPA 20k WSJ task our system obtained 11.6% word error, the 39% error on this year's evaluation suggests there are still many aspects need to be learned for a new comer like us.

1. Introduction

This paper presents the CSLU Broadcast News transcription system used in the DARPA 1997 evaluation. The system was built using software developed for the CSLU LVCSR project, initiated in January 1997. The project proceeded through development and evaluation of systems associated with previous DARPA tasks; specifically the RM system, and the WSJ-5k and WSJ-20k systems. On October 1st, 1997, work was begun on the Broadcast News task for the November, 1997 evaluation.

The 1997 Hub4 evaluation posed some new challenges to us: (1) 1997 was the first year that we started the Large Vocabulary Continuous Speech Recognition (LVCSR) project, and (2) it was the first time that we participated in a DARPA LVCSR evaluation. During the past year we spent much of our time re-inventing the wheel, and it was a good learning experience for us.

The CSLU Hub4-system is based on continuous HMMs, with a 25k-word vocabulary. WSJ SI-284 and BN training data were used for acoustic training. The decision tree-based state clustering algorithm [24] was used to cluster the phonetic contexts, which resulted in 5300 distinct states. The system was bootstrapped from our WSJ-20K system. The standard forward-backward algorithm was used for model estimation on the combined data set. The resulting cross-word triphone system has 12 Gaussians per state. The good-turing method was employed to estimate the back-off trigram language model, which resulted in a model with 10M trigrams and 5M bigrams. The pre-segment/cluster information provided by CMU was used. The evaluation system yielded 39% word error rate on the official test data.

2. A quick review of CSLU LVCSR effort

In January 1997 the Center for Spoken Language (CSLU) at CSLU assembled a three-person team and started the large vocabulary project. We used the DARPA tasks during the past 10 years as our progress milestones.

- From January to March, we worked on the Resource Management (RM) task. During this period, we implemented the basic training and decoding software.
- From April to July, we worked on the Wall Street Journal (WSJ) 5k task. During this period, we implemented the Maximum Likelihood Linear Regression (MLLR) [16, 9] and Vocal Tract Length Normalization (VTN) [15, 6], and started to play with language models.
- From August to September, we worked on the WSJ 20k system and implemented the parallel version of our training and decoding tools.

Language modeling, Maximum A Posteriori (MAP) [25, 10, 26] and Speaker Adaptive training (SAT) [2] were studied at this period.

- From October to November, we received all the training/development data related to Broadcast News (BN) and started to build the evaluation system.

Our search engine is a single pass decoder which supports word-dependent N-best results [21], high order language models, and cross-word triphone for large vocabulary continuous speech recognition. This is an extension of the token passing algorithm [23]. By decoupling the search and the search space, the tree is re-entered conceptually instead of being copied. Cross-word triphone decoding is achieved by tagging tokens differently which are being passed along the same tree nodes. Multiple pruning strategies are employed to alleviate the increased CPU cost incurred by the re-entry and the delayed application of language model probabilities.

TASK	Baseline	MLLR+VTN
RM Oct'89	3.1%	-
RM Feb'91	2.7%	-
WSJ5k ST_DT_05	10.2%	-
WSJ5k Nov'92	8.2%	5.4%
WSJ20k SI_DT_20	17.2%	-
WSJ20k Nov'92	13.7%	11.6%

Table 1: Our System Performance on the Previous DARPA Tasks

The results achieved on these tasks are summarized in Table 2., All these results are obtained using the standard training set and evaluation set. Our Resource Management system uses the standard Word-Pair grammar, and the rest use trigram models. These results compare favorably with other systems of equal complexity [24, 14, 7, 27].

3. Development of the 1997 evaluation system

At the end of September we received all the data related to the Hub4 task (acoustic and language modeling data). Due to the time constraints, all the decisions made for Hub4 specific components are based on the discussions in [3, 20, 18, 4, 8, 28, 11, 22, 13] and our understanding of these approaches.

We basically adopted BBN's strategy in the 1996 evaluation: One set of acoustic models for all the BN conditions [17]. The system was planned as:

1. Monophone recognition and acoustic wave segmentation
2. Segment clustering
3. VTN adaptation based on the decoded monophone string
4. Decode with the speaker-independent models with VTN
5. Decode with MLLR using the output from the previous step

3.1. Dictionary

The baseform of our dictionary used the LIMSI 1993 WSJ 20k word pronunciations, and we appended 5k more words to the dictionary based on word frequencies in the BN LM data. The pronunciations for these 5k words were extracted from the CMU dictionary and hand-tuned to make it "consistent" to the LIMSI dictionary. We used the same method to generate the pronunciations for words in the training data which were not in our dictionary. This dictionary has 1.6% OOV on the 1996 Hub4 evaluation set. Post evaluation analysis showed we have a 2.2% OOV rate on this year's evaluation set.

3.2. Acoustic Training

The seed models used in the Hub4 system training were from our WSJ20k system. Due to our limited CPU resources, we were never able to test all the resulting acoustic models on the complete 1996 development or evaluation data set. We randomly selected 181 segments (about 1000 seconds of speech data) from the 1996 PE data. This set was used as our development set all through the evaluation.

A number of experiments were conducted to find the viable ways to move from the WSJ task to the BN task. These include:

- Forward-Backward (FB) training with pooled WSJ (SI 284) and BN data.
- Forward-Backward training with WSJ second channel data and BN data.
- Forward-Backward training using BN data only
- SAT
- MLLR
- MAP

We were not able to make the adaptation based training method work better than the standard forward-backward training for the time being, perhaps because we did not find the optimal parameters or simply because we have bugs in our software. Some of the results are summarized in Table 3.2.

SYSTEM	WER
FB: BN only	39.1%
FB: WSJ SI284+BN	38.4%
MAP: WSJ, BN	39.6%
MLLR: BN	40.6%

Table 2: Word Error Rates (WER) for different training methods, without adaptation in decoding

3.3. Language Modeling

The CMU-Cambridge language model package V2.0 was used [19, 5]. The text materials include the WSJ LM data and BN LM data obtained from LDC. All the filler words were removed from the text and the only context cue used was the sentence begin/end. Transcriptions for the BN acoustic data were copied twice as part of the training data as suggested in [1].

The good-turing method was employed to estimate the back-off trigram language model, which resulted in a model with 10M trigrams and 5M bigrams. This language model has a perplexity of 170 on the 1996 evaluation data and a perplexity of 150 on the 1997 data (from post evaluation analysis).

3.4. Segmentation and Clustering

We experimented with the commonly adopted strategy: Use the silence segments located by a monophone recognizer as boundaries of presegments and then use some distortion measures to cluster these segments. The method proposed by [12] was implemented. We experimented with this method on concatenated WSJ utterances and found generally it worked quite well. When we experimented with the actual BN data with monophone recognition generated boundaries, we found the presegmentation generating too many very long (short) utterances. These segments could not be processed by our decoder (due to the memory requirement or adaptation requirement of duration) and also caused many cluster errors. This may be due to the fact that we trained monophones on RM data, which are acoustically quite different from BN data. We therefore decided to use the presegment/cluster information provided by CMU for this evaluation.

3.5. The Evaluation System and Results

The system used in the official evaluation is organized as follows:

1. Run the speaker-independent system on the provided segments.
2. Group all the data from the same cluster to perform MLLR and re-run the adapted system on the data in this cluster.
3. Repeat step 2 until all the clusters are processed.

Our speaker-independent acoustic model is also gender-independent (one set of models only). It is trained with WSJ SI-284 and BN training data using a standard forward-backward algorithm. The system was bootstrapped from our WSJ-20K system. The decision tree based algorithm was used to cluster the context, which resulted in 5300 distinct states. The resulting cross-word triphone system has 12 Gaussians per state (total: 63.6k Gaussians).

The results on the 1997 evaluation are summarized in Table 3.5.. When we generated these results, it was the first time our system was run on a complete data set.

SYSTEM	WER
Baseline	41.7%
Adaptation	39.0%

Table 3: Word Error Rates (WER) with/without MLLR

3.6. Resources

Our computing resources were limited to 7 Intel Pentium Pro 200 dual-CPU 512M RAM machines (3 of them were obtained at the end of August and 2 of them was obtained 2 weeks before the deadline). Our decoder runs about 300 times real time on the BN data. Due to the memory requirement of the decoder, we can only use 1 CPU per machine during decoding. The total human resources devoted to the effort during the N months of the effort was about two man years.

4. Issues

In the past year we progressed from Resource Management and Wall Street Journal Tasks, which served as our training grounds, and jumped into much more complicated BN task. Many things are still mysterious to us. There are a list of issues that we need to investigate before our system can catch up to the field. These include:

1. How to handle LVCSR recognition in noisy environments (the Hub3 type task).
2. How to handle constant changes of speaker/ environment.
3. How to handle sentence fragments due to presegmentation.
4. How to make the decoder handle OOV gracefully (on our small development set, we found roughly an OOV word caused about 4 insertion/deletion/substitution errors).
5. How to make adaptive training methods (such as MAP, SAT, MLLR) work.

5. Concluding Remarks

This paper reported our first attempt in the LVCSR research. It is a good learning experience for us. The one-year catch-up game resulted in a basic system-building software package which will serve as the research platform for our future research. We note that many of the problems that must solved require solving problems that have been encountered and solved by others before us. While many basic concepts are presented in the literature, creating a competitive system clearly involves confronting and solving many interesting problems. We hope that having to pay dues may result in some creative new ideas that will benefit the field.

6. Acknowledgement

The authors wish to thank ONR/DARPA, Fonix and Intel for the support of this work. We thank Drs. R. Stern and R. Rosenfeld at CMU, Mr. C. Che at Rugters, Mrs. F. Kubala, H. Jin, L. Nguyen and R. Schwartz at BBN, Dr. G. Adda at LIMSI, Mr. P. Clarkson at Cambridge University for answering many of our questions related to the evaluation.

7. References

[1] G. Adda, L. Lamel, M. Adda-Decker, and J.L.Gauvain. Language and lexical modeling in the limsi nov96 hub4 system. In *Proc. of DARPA Speech Recognition Workshop*, 1997.

[2] T. Anastasakos, J. McDonough, and J. Makhoul. Speaker adaptive training: a maximum likelihood approach to speaker normalization. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1043–1046, 1997.

[3] R. Bakis, S. Chen, P. Gopalakrishnan, S. Maes R. Gopinath, and L. Polymenakos. Transcription of braodcast news shows with the ibm large vocabulary speech recognition system. In *Proc. of DARPA Speech Recognition Workshop*, 1997.

[4] C. Che, D. Yuk, S. Chennoukh, and J. Flanagan. Development of the ru hub4 system. In *Proc. of DARPA Speech Recognition Workshop*, 1997.

[5] P. Clarkson and R. Rosenfeld. Statistical language modeling using the cmu-cambridge toolkit. In *Proc. of EUROSPEECH 1997*, pages 2707–2710, 1997.

[6] E. Eide and H. Gish. A parametric approach to vocal tract length normalisation. *Conference Proceedings of ICASSP'96*, I:346–348, May 7-10 1996. Atlanta, Georgia.

[7] F.Kubala, A.Anastasakos, J.Makhoul, L.Nguyen, R.Schwartz, and G.Zavaliagkos. Comparative experiments on large vocabulary speech recognition. *Proceedings of the International Conference on Acoustic Speech and Signal Processing*, 1994.

[8] F.Kubala, H. Jin, S. Matsoukas, L. Nguyen, R. Schwartz, and J. Makhoul. The 1996 bbn byblos hub4 transcription system. In *Proc. of DARPA Speech Recognition Workshop*, 1997.

[9] M.J.F. Gales and P.C. Woodland. Variance compensation within the mllr frame work. *Technical Report, CUED/F-INFENG/TR 242*, February 1996. Cambridge University, Engineering Department.

[10] J.L. Gauvain and C.H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Trans. Speech & Audio Process*, 2:291–298, 1994.

[11] G.Cook, D. Kershaw, J. Christie, and T. Robinson. The 1996 abbot broadcast news system. In *Proc. of DARPA Speech Recognition Workshop*, 1997.

[12] H.Jin, F. Kubala, and R. Schwartz. Automatic speaker clustering. In *Proc. of DARPA Speech Recognition Workshop*, 1997.

[13] J.L.Gauvain, G. Adda, L.F.Lamel, and M.Adda-Decker. Transcribing broadcast news: The limsi nov96 hub4 system. In *Proc. of DARPA Speech Recognition Workshop*, 1997.

[14] J.L.Gauvain, L.F.Lamel, G.Adda, and M.Adda-Decker. The limsi continuous speech dictation system: Evaluation on the arpa wall stree journal task. *Proceedings of the International Conference on Acoustic Speech and Signal Processing*, 1994.

[15] L. Lee and R. C. Rose. Speaker normalization using efficient frequency warping procedures. *Conference Proceedings of ICASSP'96*, I:353–356, May 7-10 1996. Atlanta, Georgia.

[16] C.J. Leggetter. Improved acoustic modelling for hmms using linear transformations. *Ph.D thesis, Cambridge University*, 1995.

[17] D. Pallett and J. Fiscus. 1996 preliminary broadcast news benchmark tests. In *DARPA 1997 speech recognition workshop*, 1997.

[18] P. Placeway, S.Chen, M. Eskenazi, U. Jain, V.Parikh, B. Raj, M. Ravishankar, R. Rosenfeld, K. Seymore, M. Siegler, R. Stern, and E. Thayer. The 1996 hub-4 sphinx-3 system. In *Proc. of DARPA Speech Recognition Workshop*, 1997.

[19] R. Rosenfeld. The cmu statistical language modeling toolkit, and its use in the 1994 arpa csr evaluation. In *ARPA Spoken Language Technology Workshop*, 1995.

[20] A. Sankar, A. Stolcks, L. Heck, and F. Weng. Sri h4-pe system overview. In *Proc. of DARPA Speech Recognition Workshop*, 1997.

[21] R. Schwartz and S. Austin. A comparison of several approximate algorithms for finding multiple (n-best) sentence hypotheses. In *ICASSP'91*, S10.4, pages 701–704, 1991.

[22] S. Sekine, A Borthwick, R. Grishman, and S. Katz. Nyu language modeling experiment for 1996 csr evaluation. In *Proc. of DARPA Speech Recognition Workshop*, 1997.

[23] S.J.Young, N.H.Russell, and J.H.S.Thornton. Token passing: A simple conceptual model for connected speech recognition systems. *Cambridge University Engineering Department Technical Report*, July 31 1989.

[24] S.J.Young and P.C.Woodland. Tree-based state-tying for high accuracy acoustic modeling. *Proc Human Language Technology Workshop*, pages pp307–312, March 1994.

[25] R.M. Stern and M.J. Lasry. Dynamic speaker adaptation for isolated letter recognition using map estimation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 734–737, 1983.

[26] E. Thelen, X. Aubert, and P. Beyerlein. Speaker adaptation in the philips system for large vocabulary continuous speech recognition. *Conference Proceedings of ICASSP'97*, pages 1035–1038, April 21-24 1997. Munich, Germany.

[27] V.Digalakis and H.Murveit. Genones: Optimizing the degree of mixture tying in a large vocabulary hidden markov model based speech recognizer. *Proceedings of the International Conference on Acoustic Speech and Signal Processing*, 1994.

[28] P. Woodland, M. Gales, D. Pye, and S. Young. Broadcast news transcription using htk. In *Proc. of DARPA Speech Recognition Workshop*, 1997.